Vocal attractiveness of statistical speech synthesisers

S. Andraszewicz, J. Yamagishi, and S. King. Vocal attractiveness of statistical speech synthesisers. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 5368-5371, May 2011.

VOCAL ATTRACTIVENESS OF STATISTICAL SPEECH SYNTHESISERS

Sandra Andraszewicz, Junichi Yamagishi, Simon King

The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

ABSTRACT

Our previous analysis of speaker-adaptive HMM-based speech synthesis methods suggested that there are two possible reasons why average voices can obtain higher subjective scores than any individual adapted voice: 1) model adaptation degrades speech quality proportionally to the distance 'moved' by the transforms, and 2) psychoacoustic effects relating to the attractiveness of the voice. This paper is a follow-on from that analysis and aims to separate these effects out. Our latest perceptual experiments focus on attractiveness, using average voices and speaker-dependent voices without model transformation, and show that using several speakers to create a voice improves smoothness (measured by Harmonics-to-Noise Ratio), reduces distance from the the average voice in the log F0-F1 space of the final voice and hence makes it more attractive at the segmental level. However, this improved attractiveness is weakened or overridden at supra-segmental or sentence levels.

Index Terms— average voice, attractiveness, speaker adaptation, speech synthesis, HMM

1. INTRODUCTION

Vocal attractiveness has been examined in various fields. [1] mentions that an employee's voice attractiveness affects customer's attribution to that employee's expected and perceived effort and ability and the customer's expected and perceived encounter satisfaction. The author refers to the classic phenomenon, first coined by [2], of "What is beautiful is good." Later this theory was extended by [3] into the form "What sounds beautiful is good". It has been shown that attractive voices have the same effect as attractive faces, meaning that vocal attractiveness parallels visual attractiveness. Hence, [4] investigated the acoustic correlates that make a voice attractive. They found that attractive voices were louder and more resonant, but also more intermediate in loudness and resonance. In addition to this, they found some sex differences, including that low-pitch-male voices are perceived as more attractive, while the attractiveness of female voices could not be captured by spectrographic analysis.

As mentioned above, voices with intermediate features were found more attractive. [5] created averaged voices out of 2, 4, 8, 16 and 32 composite male or female voices using STRAIGHT interpolation functions [6] and found that the more voices are averaged, the more attractive the averaged voice is perceived. Further, they found that the attractiveness scores are correlated with both vocal smoothness measured by Harmonics-to-Noise Ratio (HNR) and with distance from the average voice in the log F0-F1 space of the final voice; these are additive. In their study, the stimuli were created from the pre-recorded word "had" uttered by 64 speakers. As a stimulus for judging the attractiveness of a voice, they used the phone /A/ excised from the word "had" and averaged it across various numbers of speakers. Since it was not clear to us whether the "average voice is more beautiful" effect would hold for speech generated from statistical models, or for longer utterances, such as words or sentences, we decided to test this hypothesis.

In text-to-speech synthesis, the average voice plays an important role as the basis for speaker-adaptive HMM-based speech synthesis where a speaker's characteristics can be "cloned" using a small amount of their speech data [7] and, interestingly, our previous perceptual experiments using an average voice and 60 cloned voices [8] showed very similar tendencies to [5]; that is, the average voice typically obtains a high subjective score for *naturalness* and the scores for cloned voices are correlated with distance from the average voice.

There are at least two possible conclusions that could be drawn from this result: either the model transformation techniques used for the voice cloning simply reduce naturalness as the distance moved from the average voice increases, or our subjects were rating something closer to *attractiveness* than *naturalness*.

Therefore, the first goal of the current work is to check whether the effect found by [5] – the more speakers that are averaged, the more attractive the result is – applies not only to speech created using the STRAIGHT interpolation functions [6] but also to speech generated from statistically-averaged models vocoded using STRAIGHT. This effect would be mediated by improvements in "smoothness" caused by averaging the voices and by decreasing the distance of both F0 and the first formant from the voice population mean. The second goal is to extend the experiments to longer utterances such as words and sentences. For these goals we adopted the same *rating task* and experimental designs for measuring attractiveness as used by [5] and performed very similar perceptual experiments, except using average voices generated from speaker-independent HMM-based speech synthesisers and individual speaker's voices generated from speaker-dependent HMM-based speech synthesisers.

2. EXPERIMENTAL CONDITIONS

2.1. Experimental Design

This experiment consisted of one task and a debriefing questionnaire. Each participant was presented with 240 stimuli (80 phonemes, 80 non-words and 80 sentences), divided into 3 equal blocks of trials. Contrary to the method of [5], in the current experiment, the number of stimuli of each type was the same and equalled 10.

20 postgraduate students and research fellows of the University of Edinburgh (3 female, 17 male, age:22-37) participated in the experiment. Based on the results from preliminary experiments, it was assumed that the gender of subjects does not influence the rating of attractiveness of the voices. All participants had normal or corrected-to-normal vision, reported no major hearing problems and had a good command of English.

2.2. Materials

The Wall Street Journal 1 corpus [9] was used to create the stimuli. We used both the long training data (1200 sentences per speaker)

and short training data (150 sentences per speaker). In order to produce statistical models in which the speech of 1, 2, 4, or 8 speakers was averaged, only the speakers with long training data were used, while for producing 16, 24, 32, and 64-speaker voices, speakers with both, long and short training data were used. In the latter case, maximum of 12 speakers with long training data was used and the remaining speakers were with short training data, e.g. 16-speaker voices was included: 12 long-data-speakers and 4 short-data-speakers.

Because of the large number of possible combinations of speaker which one could use to create the multiple-speaker average voices, the choice of the speakers and the sentences from the corpus that were used was randomised. The main difference between single and multiple-speaker voices was that for the single-speaker voices the corpus of only one speaker was used, while for the multiple-speaker voices the corpus was equally divided across the speakers. The number of sentences used to train all voices was the same: 1200 sentences. We created 8 types of models: single-speaker, 2-speaker, 4-speaker, 8-speaker, 16-speaker, 24-speaker, 32-speaker, 64-speaker models. There were 12 versions of each type, which in total resulted in 96 different voices. For simplicity, the experiment only used male models.

Speaker-independent HMMs for each voice (speaker-dependent, in the case of the 1-speaker voices) were trained using the HTS-2008 framework [8] with Festival text-processing. For each of the 96 voices, a phoneme, a non-word and a sentence were synthesised. Thus, each voice uttered the following:

1. /a/
2. *flane*
3. *The television is in the living room.*

The phoneme stimuli were first cut from a carrier word /a a/. Contrary to the study of [5] phoneme /a/ rather than phoneme /A/ was chosen, because the latter one when synthesised sounded very breathy for some of the voices, while /a/ was well synthesised and clear for all the voices. The new word consisted of a repeated phoneme because when synthesising a single phoneme, for some voices one could perceive 'a transition to another vowel', which presumably resulted from the training data. Synthesising a single 'word' from two of the same vowels reduced the effect of formant transitions and influence of the vowel environment in the training data. In the new word, a very short break between the two phonemes was observed and there was no significant formant transition.

From the synthesised word composed of two /a/ phonemes ,the second phoneme and the silence following it were removed. The resulting waveform was noise-gated and 250 msec silence was added at the end of the sound. The resulting stimuli were somewhat shorter than the stimuli used by [5], which were 201—477 msec. Although it would have been possible to use PSOLA or some other method to extend the duration of each phoneme stimulus to match the durations reported by [5], this may have significantly increased the Harmonics-to-Noise Ratio, so we decided not to do this.

Phoneme, non-word and sentence stimuli were synthesised for all 96 voices. However, some of the voices sound very buzzy. As a consequence, only 10 versions of each type of model were retained. The selection was based on the measurement of Harmonics-to-Noise Ratio. Therefore in total there were 240 stimuli, where 80 were *phonemes*, 80 were *non-words* and 80 were *sentences*.

2.3. Procedure

The experiment was divided into three blocks of 80 stimuli. For each participant the order of the stimuli was randomised and the stimuli of each length were mixed up. Attractiveness rating was done on a 5-point Lickert scale and the 3 blocks of trials were followed by a debriefing questionnaire. Each participant was placed in a separate soundproof cubicle of a perception lab located at the University of Edinburgh and listened to the stimuli through good quality headphones. The experiment took participants an average of about 40 minutes to complete.

3. ANALYSIS AND RESULTS

3.1. Attractiveness of Voices

As in [5], raw scores of attractiveness measured on a 5-point scale were first normalised per participant. Mean normalised scores were calculated for each stimulus and are presented in Figures 1–3. The figures show that, in general, averaging the voices results in narrowing the range of the attractiveness scores. Also, it can be seen that there are differences in attractiveness of the voices depending on the length of the utterance. Indeed, a one-way-ANOVA indicated significant differences between the stimuli of different length ($F(2, 237) = 75.07, p < .001, R^2 = .39$). However, the variances of scores of stimuli of each length were not homogeneous ($F(2, 237) = 6.96, p = .001$). According to a Bonferroni post-hoc test, phonemes ($Mean = .35, SD = .07, SE = .01$) were perceived as less attractive than non-words ($Mean = .48, SD = .11, SE = .01$), $p < .001$, and than sentences ($Mean = .52, SD = .10, SE = .01$), $p < .001$. Non-words were perceived as less attractive than sentences, $p < .01$.

In addition to this, the scores were higher for the averaged voices than for the single-speaker voices. In the current study there was always the same number of versions for each, whereas [5] had fewer versions for the stimuli consisting of more speakers. A significant improvement in attractiveness can be observed as the number of speakers used to train a model increases. Single-speaker voices uttering the phoneme and the non-word were judged as the least attractive over all models, while in case of sentences, single-speaker voices were judged as the third least attractive. Importantly, the most attractive models in the case of phoneme-stimuli were those consisting of 64 speakers. In the case of non-words, 64-speaker voices were only 6th most attractive, and in the case of the sentences, the second most attractive. The most attractive voices uttering the phoneme were 2-speaker voices, and for the sentence the most attractive were 4-speaker models.

Three One-Way-ANOVAs were conducted to investigate the differences between the mean attractiveness scores of the 8 types of models, separately for phoneme, non-word and sentence stimuli. All three tests indicated significant differences among different models ($F(7, 72) = 4.70, p < .001, R^2 = .31$, $F(7, 72) = 5.48, p < .001, R^2 = .40$, and $F(7, 72) = 5.56, p < .001. R^2 = .35$ consecutively). In the case of phonemes, a post-hoc LSD test indicated that single-speaker models were significantly less attractive than all multi-speaker models ($p < .01$), and 4-speaker models were significantly less attractive than 64-speaker models ($p < .05$); there were no significant differences among other multi-speaker models. In case of non-words, single speaker models were significantly less attractive than all models except for 4-speaker and 32-speaker models, which were less attractive than the single-speaker models ($p < .05$) and 64-speaker models were only more attractive than the single-speaker models. The most attractive 2-speaker models obtained higher scores than all other models apart from 16-speaker models ($p < .05$). For the sentence-stimuli, single-speaker models were significantly less attractive only than 4-speaker models ($p = .001$), which were the most attractive models. In contrast, the

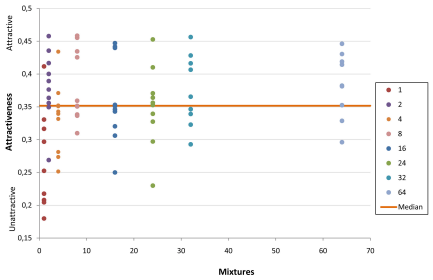

Fig. 1. Phoneme stimuli

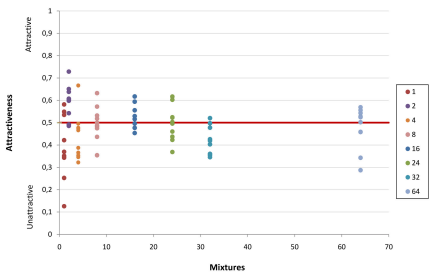

Fig. 2. Non-word-stimuli

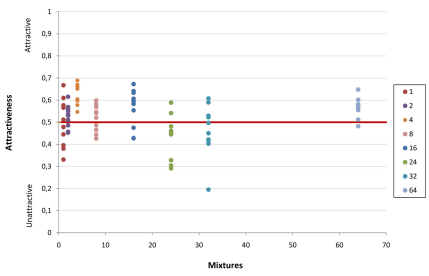

Fig. 3. Sentence stimuli

(a) Z scores of voice attractiveness of each voice. Horizontal axis shows the number of speakers used for building voice.

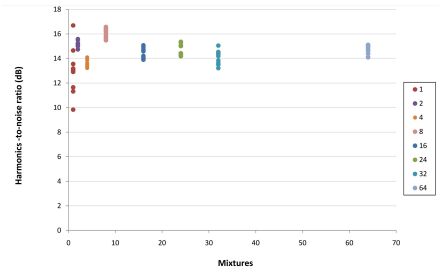

Fig. 4. Phoneme stimuli

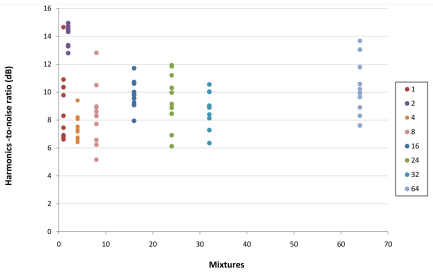

Fig. 5. Non-word-stimuli

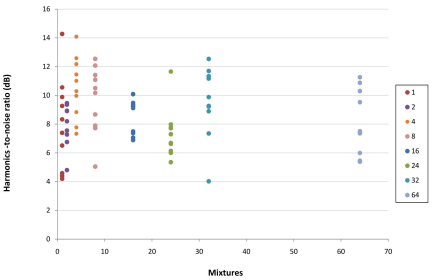

Fig. 6. Sentence stimuli

(b) Harmonic-to-noise ratio of each voice. Horizontal axis shows the number of speakers used for building voice.

4-speaker models were more attractive than all other models apart from 16- and 64-speaker models. The variances of phoneme and non-word scores were homogeneous, while variances of the scores of the sentences were inhomogeneous.

3.2. Influence of Harmonics-to-Noise Ratio on the Voice Attractiveness

The Harmonics-to-Noise Ratio (HNR) represents the "harshness" and "smoothness" of voices [10]. [5] found that HNR is correlated with voice attractiveness. Figures 4 – 6 shows the HNR of our voices and we can see that averaging results in narrowing the range of HNR and generally improving HNR of the synthesised voices. These two trends are especially pronounced for the phoneme stimuli.

Three One-Way-ANOVA tests showed significant differences in HNR among different models, for phoneme- ($F(7,72) = 15.19p < .001, R^2 = .60$), non-word- ($F(7,72) = 13.82, p < .001, R^2 = .57$) and sentence-stimuli ($F(7,72) = 2.494, p < .05, R^2 = .20$). However, HNR variances of the phoneme- and non-word stimuli were inhomogeneous ($F(7,72) = 4.59, p < .001$, and $F(7,72) = 2.84, p < .05$, respectively). LSD post-hoc analyses revealed that HNRs of single-speaker voices uttering phonemes were significantly lower than all other models apart from 4-speaker model ($p = .001$), while 8-speaker models have higher HNRs than all other models ($p < .05$). Similarly, in the case of non-words, 2-speaker models had significantly higher HNR than all other models ($p < .001$), while HNRs of 4-speaker models were significantly lower than all other models apart from 8- and 32-speaker models ($p < .05$). Finally, in the case of the sentence the models differed in HNR, but no clear-cut pattern could be observed.

Figures 7 – 9 show Pearson's correlations between HNR and attractiveness scores. There was a moderate positive correlation for phonemes, $r = .428, p < .001$, moderate positive correlation for non-words, $r = .396, p < .001$, and no correlation between these variables for sentences, $r = .036, p > .05$.

3.3. Influence of Distance from the logF0-logF1 mean on the Voice Attractiveness

[5] mentioned that voice attractiveness is also correlated with distance from the logF0-logF1 mean. As shown in Figures 10 – 12 there was a moderate negative correlation (Pearson) between distance from the logF0-logF1 mean and attractiveness scores for phonemes, $r = -.246, p < .05$, no correlation was found for the non-words and sentences, $r = .138, p > .05$, and no correlation between these variables for sentences, $r = -.114, p > .05$.

4. CONCLUSION

Our perceptual experiments can be summarised as follows: Statistically averaging speakers results in more attractive voices, confirming the findings from [5]. In general, this is true not only at the phoneme level but also at word or sentence levels. At phoneme levels, HNR and distance from the logF0-logF1 mean were found to be correlated with voice attractiveness in a similar way to [5]. However, these measures could not explain voice attractiveness measured using words or sentences. Future work is therefore needed to explore measures that can explain voice attractiveness at the supra-segmental levels. An extended report of the work reported in this paper can be found in [11].

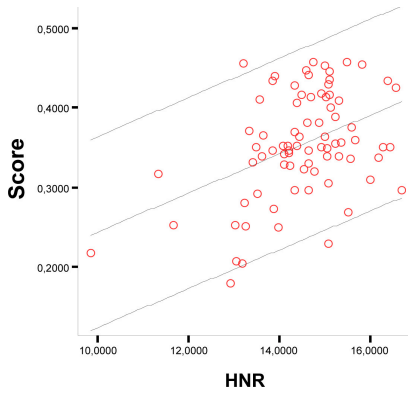

Fig. 7. Phoneme stimuli

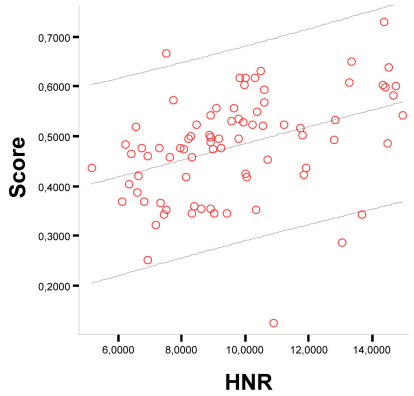

Fig. 8. Non-word-stimuli

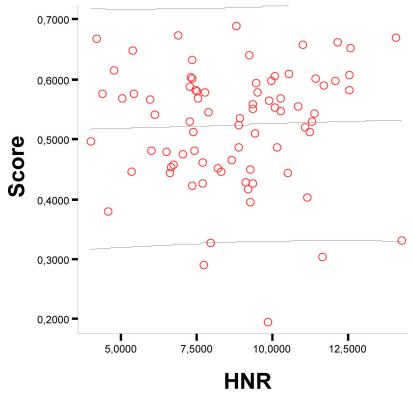

Fig. 9. Sentence stimuli

(a) Correlation between Z scores of voice attractiveness and harmonics-to-noise ratio

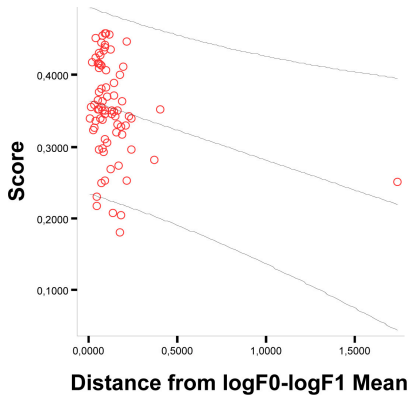

Fig. 10. Phoneme stimuli

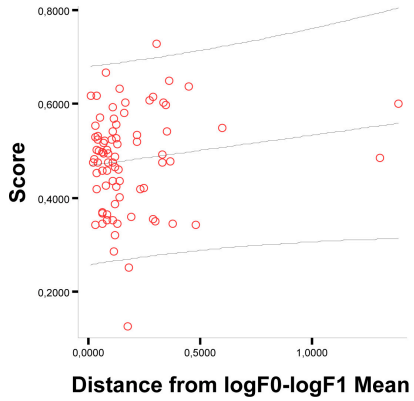

Fig. 11. Non-word-stimuli

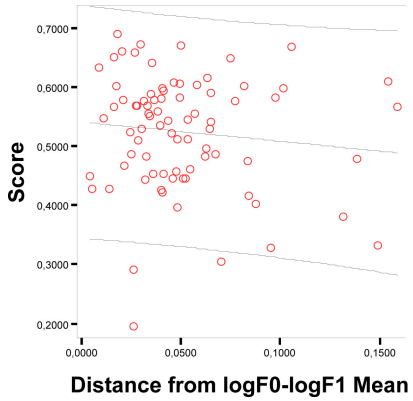

Fig. 12. Sentence stimuli

(b) Correlation between Z scores of voice attractiveness and logF0-F1 mean

5. REFERENCES

[1] S. Bartsch, ""What sounds beautiful is good?" How employee vocal attractiveness affects customers evaluation of the voice-to-voice service encounter," *Aktuelle Forschungen im Dienstungsmarketing*, pp. 45–68, 2009.

[2] K.K. Dion, E. Berscheid, and E. Walster, "What is beautiful is good," *Journal of Personality and Social Psychology*, vol. 24, pp. 285–290, 1972.

[3] M. Zuckerman and R.E. Driver, "What sounds beautiful is good: The vocal attractiveness stereotype," *Journal of Nonverbal Behavior*, vol. 13, no. 2, pp. 67–82, 1989.

[4] M. Zuckerman and K. Miyake, "The attractive voice: What makes it so?," *Journal of Nonverbal Behavior*, vol. 17, no. 2, pp. 119–135, 1993.

[5] L. Bruckert, P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G.A. Rousselet, H. Kawahara, and P. Belin, "Vocal attractiveness increases by averaging," *Current Biology*, vol. 20, pp. 116–120, 2009.

[6] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[7] J. Yamagishi, *Average-Voice-Based Speech Synthesis*, Ph.D. thesis, Tokyo Institute of Technology, 2006.

[8] Junichi Yamagishi, Oliver Watts, Simon King, and Bela Usabaev, "Roles of the average voice in speaker-adaptive HMM-based speech synthesis," in *Proc. Interspeech 2010*, Sept. 2010, pp. 418–421.

[9] Linguistic Data Consortium, "Wall Street Journal-based continuous speech recognition (CSR) corpus phase II (WSJ1): Training and development test texts and documentation," Tech. Rep., University of Pennsylvania, April 1994.

[10] C.T. Ferrand, "Harmonics-to-noise ratio: An index of vocal aging," *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[11] Sandra Andraszewicz, "Are synthetic average voices more beautiful?," M.S. thesis, The University of Edinburgh, 2010.